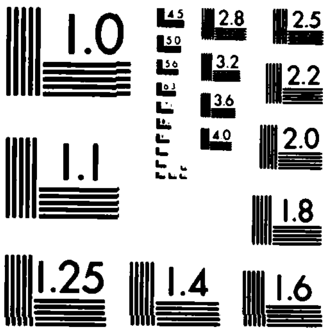
AD-A150 065   A NOTE ON VALIDITY GENERALIZATION PROCEDURES(U) GEORGIA   1/1
INST OF TECH ATLANTA SCHOOL OF PSYCHOLOGY
L R JAMES ET AL. JAN 85 GT-ONR-7 N00014-83-K-0480

UNCLASSIFIED   F/G 12/1   NL

END

FILMED

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
GT-ONR-7

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
A Note on Validity Generalization Procedures

5. TYPE OF REPORT & PERIOD COVERED
Interim Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Lawrence R. James, Robert G. Demaree, and Stanley A. Mulaik

8. CONTRACT OR GRANT NUMBER(s)
N00014-83-K-0480

9. PERFORMING ORGANIZATION NAME AND ADDRESS
School of Psychology
Georgia Institute of Technology
Atlanta, Georgia 30332

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR475-026

11. CONTROLLING OFFICE NAME AND ADDRESS
Manpower R & D Program
Office of Naval Research (Code 270)
Arlington, Virginia 22217

12. REPORT DATE
January, 1985

13. NUMBER OF PAGES
46

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)
Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Cross-Situational Consistency
Cross-Situational Specificity
Validity
Validity Generalization

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Validity generalization procedures are reviewed and found to be subject to the logical fallacy of affirming the consequent. Alternative models may explain variation in validity coefficients as well as the cross-situational consistency model espoused by proponents of validity generalization. Moreover, many of the assumptions that form the statistical foundation for validity generalization are false. It is recommended that (a) the conditionality of inferences based on validity generalization analyses be

85 01 30

20.

explicitly stated, (b) the decision rule for the validity generalization ratio be changed from .75 to .90, and (c) validity generalization analyses employ Fisher z coefficients rather than (Pearson) correlation coefficients.

A Note on Validity Generalization Procedures

We have been informed recently that the *scientific status* of personnel research will "greatly advance" if the hypothesis that validities are situationally specific is found to be false (Schmidt, Hunter, & Pearlman, 1982, p. 841; see also Schmidt & Hunter, 1977, 1978, 1980). The term "situational specificity" holds that true validities (i.e., population correlations unaffected by statistical artifacts, indicated by p_i, i = 1,...,K populations) vary as a function of validation situation (setting, study), or $\sigma_p^2 > 0$ (cf. Hunter, Schmidt, & Jackson, 1982). In contrast to situational specificity, we will use the term "cross-situational consistency" to refer to conditions in which the p_i are constant over K populations (situations) and all variation among observed validities (r_i, i = 1,...,K) is attributable to sampling error and other types of statistical artifacts such as variations in criterion reliabilities and variances (cf. Hunter & Hunter, 1984; Hunter, Schmidt & Jackson, 1982; Hunter, Schmidt, & Pearlman, 1981, 1982; Schmidt, Hunter, Pearlman, & Shane, 1979; see also Callender & Osburn, 1980, 1981, 1982; Raju & Burke, 1983). A constant p across situations implies that validity is generalizable or "transportable" (Schmidt et al., 1982, p. 81), although the term "validity generalization" refers to a less demanding condition in which "most of the values of estimated true validities...lie in the positive range" (Schmidt et al., 1982, pp. 840, 841).

The term "validity generalization approach (analysis, procedure)" is employed here to refer broadly to the assumptions and quantitative techniques used by the proponents of this approach to contrast cross-situational consistency with situational specificity. It is noteworthy that proponents of the validity generalization approach have based their substantive and statistical developments on *structural equations* for observed validities, the objective being to identify *causes* for variation among validity coefficients and thereby construct *explanatory models* for validity distributions (cf. Schmidt, Gast-Rosenberg, & Hunter, 1980). Indeed, the attempt to construct useful explanatory models for validity distributions is made possible because structural equations provide explicit, quantitative statements of statistical theory regarding the rules that presumably govern the occurrences of validities. The validity generalization approach also proposes quantitative methods for assessing the goodness of fit of the structural equations--that is, for confirming or disconfirming predictions evolving from the causal models for validities and validity distributions. The end-products of these tests have serious implications for industrial-organizational psychologists, an example being that confirmation of a cross-situational consistency model implies that validity studies may not have to be repeated in each situation in which a test is used.

Industrial-organizational psychologists are playing for high stakes here, and rigorous review is needed of the statistical foundations for the structural equations (i.e., causal models) for validities and validity distributions, the methods for confirming or disconfirming predictions evolving from the causal models, and the causal inferences that derive from results of the confirmatory (i.e., validity generalization) analyses. Our objective is to furnish at least a partial review. To set the stage for the review, consider the following quotations, which describe what the key proponents of the validity generalization approach intended to do and their perceptions of the results of their efforts.

> In order to establish such patterns of relationships, it is first necessary to demonstrate that the doctrine of situational specificity is false or essentially false. If the situational specificity hypothesis is rejected, then it follows that various constructs--for example, spatial ability--have *invariant population relationships* with specified kinds of performances and job behaviors (Schmidt et al., 1979, p. 267, italics added).

> Schmidt and Hunter (1977) showed that ignoring sampling error leads to disastrous results in the area of personnel selection. Because he ignored the effect of sampling error on the variance of findings across studies, Ghiselli (1966, 1973) concluded that tests are only valid on a sporadic basis, that validity varies from one setting to another because of subtle differences in job requirements that have not yet been discovered (Hunter & Hunter, 1984, p. 77).

> In conclusion, our evidence shows that the validity of the cognitive tests studied is neither specific to situations nor specific to jobs (Schmidt & Hunter, 1981, p. 1133).

> The evidence from these two studies appears to be the last nail required for the coffin of the situational specificity hypothesis (Schmidt,

Hunter, Pearlman, & Hirsh, 1984, p. 73).

The problem with these conclusions is that they are stated in a categorical manner that implies irrefutable evidence in favor of a cross-situational consistency structural model and against a situational specificity structural model. Little attention is given to the possibility that future tests based on different assumptions might disconfirm cross-situational consistency or at least furnish an alternative view that explains the data as well as cross-situational consistency. To be specific, empirical support for a causal theory implies that a theory is a useful guide to explanation. It does not imply that the theory is true or unique because (a) empirical analyses usually involve untested assumptions that may be false, and (b) a set of observed data may be explained equally well by more than one causal theory (James, Mulaik, & Brett, 1982). These would seem to be pertinent concerns given that the null hypothesis of situational specificity had not been rejected in 54% (80 of 151) of the validity distributions reviewed by Schmidt, Hunter, and colleagues at the time of the Schmidt et al. (1982) article.

Our objective is to demonstrate that alternative assumptions and views exist even though validity generalization procedures appear to support a causal inference that validities are cross-situationally consistent. The first step toward this objective is to use validity generalization procedures to show that a cross-situational consistency causal model has a good empirical fit with a contrived distribution of observed validities. A number of simplifying assumptions were made with respect to the validity distribution and the analyses in order to focus on matters of principle. The simplifying assumptions were: (a) sampling error explains the lion's share of variation among observed validities (Hunter & Hunter, 1984; Schmidt et al., 1982) and thus sampling error is the only statistical artifact introduced into the distribution; (b) the sample size for each sample ($\underline{n}_{\underline{i}}$) is a constant equal to 70, which simplifies equations but retains reasonable and realistic sampling error (Lent, Aurbach, & Levin, 1971); (c) only one sample was obtained from each of $\underline{K}$ situations, which is the typical case in practice; and (d) sampling was done randomly from a bivariate normal population underlying each situation.

The second step toward the goal of demonstrating alternative views and assumptions involves proposing an alternative explanatory model to cross-situational consistency and then showing that this alternative model also has a good empirical fit with the (same) contrived distribution of observed validities. The third and final step is to show that many of the statistical assumptions on which validity generalization analyses are based are false. The paper is concluded with recommendations for future uses of validity generalization procedures and research on cross-situational consistency versus situational specificity.

An Overview of the Validity Generalization Approach

The validity generalization approach is a form of a statistical "what if" scenario. One devises a statistical scenario, applies the scenario to data on the assumption that the scenario is valid, and ascertains the statistical consequences. The key "what if" assumption for the validity generalization procedure is: What if "...the population correlation is assumed to be constant over studies" (Hunter, Schmidt, & Jackson, 1982, p. 40)? Let's proceed as if this assumption is valid for the 30 contrived validity coefficients (observed correlations or r_i, i =1,..., 30 studies, situations, populations) presented in Table 1. The distribution of contrived validities was based on the premises that (a) the "true validity" for tests for many jobs is at least .50 (Hunter & Hunter, 1984; Pearlman, Schmidt, & Hunter, 1980); (b) sampling error is the only statistical artifact in operation; and (c) the sampling distribution has a slight negative skew (for reasons addressed later). In addition, the distribution was purposefully designed to illustrate a condition in which multiple conclusions could be drawn regarding causes of variance among the r_i, the supposition being that empirical confirmation of more than one explanatory model precludes exclusive reliance on a particular explanatory model (e.g., cross-situational consistency). In this regard, the range of correlations in the contrived distribution is about the same as the range of simulated true validities used by Osburn, Callender, Greener, and Ashworth (1983, p. 117) in their "moderate true variance" distribution. The original (and simplified) validity generalization (VG) equations based on Hunter, Schmidt, and Jackson (1982), and their ensuing estimates for the data in Table 1, are presented in Table 2.

Insert Tables 1 and 2 about here

Equation 1 in Table 2 furnishes the mean observed validity ($\bar{r}$), which is an estimate of the constant population correlation "p" if indeed a constant correlation is a viable alternative to the null hypothesis of situation specificity (i.e., $\sigma_p^2 > 0$). The null hypothesis is tested by comparing the variance among the observed validities (i.e., s_r^2, Equation 2) to an estimate of the variance among these validities that would be expected from sampling error exclusively (i.e., $\hat{\sigma}_e^2$, Equation 3). The comparison typically takes the form of the ratio $\hat{\sigma}_e^2/s_r^2$, which is:

> the proportion of observed variance (the denominator) that is accounted for by statistical artifacts (the numerator). The numerator in this ratio is the variance in observed validities predicted from artifacts alone; the denominator is the observed (computed) variance of these validities. We have used this ratio to draw conclusions about the situational specificity hypothesis, that is, the hypothesis that $[\sigma_p^2] > 0$. The rule that we have used in our research is that if this ratio (expressed as a percentage) is 75% or greater, we reject the hypothesis that $[\sigma_p^2] > 0$. The rationale for this decision rule is that the remaining artifacts for which we cannot correct are likely to account for at least 25% of the observed variance (Schmidt et al., 1982, p. 840; terms in brackets reflect statistical designations used in the present discussion).

The ratio $\hat{\sigma}_e^2/s_r^2$ reported in Table 2 is .75 (75%), which satisfies the VG decision rule. According to this rule, the conclusion should be that the situational specificity hypothesis ($\underline{\sigma}_{\underline{p}}^2 > 0$) is disconfirmed because essentially 100% of the variation among validities in Table 1 can be attributed to sampling error and other, unmeasured statistical artifacts. Moreover, according to the Schmidt et al. (1979, p. 267) rationale quoted earlier, it follows from rejection of the situation specificity hypothesis that the predictor construct (e.g., spatial ability) has "invariant population relationships with specified kinds of performances and job behaviors." This is cross-situational consistency.

Affirming the consequent. Does it in fact follow that population relationships are invariant if the situational specificity hypothesis is rejected? The answer is no. Indeed, the Schmidt et al. (1979) statement illustrates a form of logical fallacy known as "affirming the consequent" (cf. James et al., 1982). This logical fallacy occurs when a good fit between predictions from a causal theory and empirical data is used to infer that the theory actually and uniquely explains the data. The fallacy of such an inference is, as noted, that other causal theories may explain the same data as well as the theory of interest and that assumptions used to conduct the empirical test may be false. To avoid the fallacy of affirming the consequent, one notes that (a) empirical support for a causal theory indicates that the theory furnishes a useful basis for explanation without (b) inferring that the theory furnishes a unique basis for explanation.

In the present case, the finding that $\hat{\underline{\sigma}}_{\underline{e}}^2/\underline{s}_{\underline{r}}^2 = .75$ indicates a good empirical fit between the data and a causal theory (explanatory model) of cross-situational consistency--according to the VG decision rule, that is. Accepting the VG decision rule as valid for the moment, the inference should be that cross-situational consistency furnishes a useful explanation for the observed variance among the $\underline{r}_{\underline{i}}$. The inference should not be that the population correlations are invariant because this implies that cross-situational consistency furnishes the only, or a unique, explanation for the data. Indeed, an exclusive attribution to cross-situational consistency is an illustration of affirming the consequent because alternative views (explanations) are consistent with the data in Table 1 and because untested assumptions can be shown to be false. The issue of alternative explanations is addressed below. This discussion is followed by consideration of false assumptions.

Alternative Explanations

In the interest of constrast to the VG approach, it is assumed now that Ghiselli (1966, 1973) was correct in concluding that validity is situationally specific. Situational specificity is presumed to be due in part to unknown differences in the measurement (latent structure) models for job performance and in job requirements over studies (situations) (Ghiselli, 1966, 1973). It is presumed further that situational specificity among correlations between a person variable predictor (e.g., cognitive skills) and a criterion (e.g., job performance) is also a potential function of moderating effects due to variation among situations in variables such as leadership, reward structures and processes, group cohesiveness, stress and coping mechanisms for stress, systems norms and values (e.g., conformity, loyalty), socialization strategies, formal and informal communication nets, formalization and standardization of structure, and physical environments (e.g., privacy), to name a few variables. While we presently lack empirically confirmed, explanatory models for job performance that integrate situational variables and person variables (cf. James & Jones, 1976), there is no dearth of basic psychological theories that portray behavior (which includes job performance) as a function of person variables, situational variables, and various forms of interactions between person variables and situational variables, including nonadditive person by situation interactions (cf. Bowers, 1973; Ekehammer, 1974; Endler & Magnusson, 1976; Lewin, 1938; Lichtman & Hunt, 1971; Pervin, 1968). It is a simple matter to employ these theories to develop models in which the correlation between a person variable and job performance varies as a function of levels or scores on situational variables. Furthermore, if we postulate that no two situations have an identical pattern of scores on the situational variables (moderators), then we may logically entertain the notion that each situation represents a different (sub)population with a different (sub)population validity.

We will therefore proceed to implement a situational specificity "what if" scenario based on the assumption that a unique population validity underlies each situation (study). We begin with the psychometric analogy employed by Hunter, Schmidt, and Jackson (1982) to establish a statistical foundation for VG analysis. The basic structural (causal) equation is:

$$\underline{r}_{\underline{i}} = \underline{p}_{\underline{i}} + \underline{e}_{\underline{i}} \qquad (5)$$

In terms of the psychometric analogy, r_i is the observed score (correlation) for a subject (sample) from population (situation) i, p_i is the true score (population correlation) for situation i, and e_i is the random measurement (sampling) error associated with r_i. The situational specificity hypothesis is again $\sigma_p^2 > 0$. Here, however, we have as many populations as we have situations or studies, which denotes that each of the 30 observed correlations in Table 1 is a single representation of its specific p_i. That is, only one random sample ($n_i = 70$) has been drawn from the specific bivariate normal distribution associated with each situation. We may also view each r_i as a single realization (sample of one) from a sampling distribution comprised of an infinite number of independently estimated correlations, where each correlation is based on a sample of 70 subjects drawn randomly from a population having p_i as a correlation. There are 30 such sampling distributions.

A point that is typically not considered in VG analysis is that "reasonable limits" for each p_i can be estimated based on the inequality $r_i - 2(\hat{\sigma}_{ei}) < p_i < r_i + 2(\hat{\sigma}_{ei})$, where $\hat{\sigma}_{ei}$ is an estimate of the error of measurement for population i (Gulliksen, 1950, p. 20). The equation for $\hat{\sigma}_{ei}$ is addressed later in this paper. To illustrate the use of reasonable limits, $\hat{\sigma}_{ei}$ for $r_i = .26$ is .11 and reasonable limits for the true (population) correlation are .04 to .48. An estimate of $\hat{\sigma}_{ei}$ for $r_i = .72$ is .06, and the reasonable limits for p_i are .60 to .84. If we were to establish reasonable limits for each of the 30 p_i and then view the 30 ranges jointly, we would find that the joint range of reasonable limits of possible values of the 30 $\underline{p}_{\underline{i}}$ varies from .04 to .84.

We now have two interpretations for the same set of correlations, one furnished by VG analysis that suggests that the $\underline{p}_{\underline{i}}$ are constant and equal to .50, and one furnished by a situational specificity hypothesis that suggests that the $\underline{p}_{\underline{i}}$ are different and could range between .04 and .84. The VG analysis reported in Table 2 supports the former, cross-situational consistency model. What evidence is there for the latter, situational specificity model? Well, we have an analysis based on chi-square to test the null hypothesis that $\underline{p}_{\underline{i}} = \underline{p}$ for all $\underline{i}$, or $\sigma_{\underline{p}}^2 = 0$. This test, given in Cohen and Cohen (1975, p. 52; the equation in Cohen & Cohen, 1983, p. 55 is missing a salient bracket), furnishes a chi-square value of 42.009, which is significant at the .05 level using a one-tail test of significance.

Rejection of the null hypothesis of cross-situational consistency implies that $\sigma_{\underline{p}}^2 > 0$, or that the validities in Table 1 may be situationally specific. We must be careful not to affirm our own consequent, however, and thus we conclude that the results of the present analysis imply that the distribution of observed validities in Table 1 could have been generated by a set of different $\underline{p}_{\underline{i}}$, plus sampling error. We have _no proof_ that this is so, but we do have a viable, empirically confirmed alternative to the assumption that the observed validities were generated by a constant $\underline{p}$ and sampling error. It may be discomforting to realize that $\sigma_{\underline{p}}^2 = 0$ and $\sigma_{\underline{p}}^2 > 0$ are both viable alternatives. Yet, multiple and conflicting explanatory models are to be expected in causal, or confirmatory, analysis (James et al., 1982). When confronted with conflicting models, the objective is to ascertain if one or more of the models might be disconfirmed by additional tests, a source of which is further examination of untested assumptions of one or more of the models. Presented below is an examination of the assumptions underlying the VG decision rule that $\sigma_p^2 > 0$ should be rejected when $\hat{\sigma}_e^2/s_r^2 \geq .75$.

A comparative analysis of power. Hunter, Schmidt, and Jackson (1982, p. 47) did not endorse their form of a chi-square test, which furnishes a chi-square value of 41.07 for the data in Table 1, because the chi-square test "has very high statistical power and will therefore reject the null hypothesis [of cross-situational consistency] given a trivial amount of variation across studies." In the interest of fairness, we believe that it is important also to evaluate the power of the VG ratio $\hat{\sigma}_e^2/s_r^2$ in regard to rejecting the null hypothesis of situational specificity. A recent simulation study by Osburn et al. (1983) suggested that the decision rule to reject the situational specificity hypothesis when $\hat{\sigma}_e^2/s_r^2 \geq .75$ results in too much power in the sense that situational specificity is rejected when low to moderate variance exists among the ρ_i (true validities), given that the n_i are not large (< 100). We wish to address this point with some logic and simple algebra within the context that sample sizes are not large (e.g., $n_i = 70$) and for the critical value of the VG decision rule (i.e., $\geq .75$).

Consider the following statement by Schmidt et al. (1982, p. 844): "We have found that, except when study sample sizes are very large, most of the variance in observed correlations that is due to artifacts is due to only one artifact--simple sampling error." To illustrate this point, Schmidt et al. (1982) reported that an average of 90% of all of the variance due to measurable artifacts was attributable to sampling error in two studies that employed the Schmidt and Hunter "interactive equation," which uses a simultaneous procedure to estimate variance due to measurable artifacts. Measurable artifacts included between-study differences in sampling error, range restriction, criterion reliability, and predictor reliability. These points suggest that for a VG ratio equal to the critical value of .75, we would attribute 67.5% [i.e., .90(.75)100] of the total observed variance (i.e., $\underline{s}_{\underline{r}}^2$) to sampling error and 7.5% [i.e., .10(.75)100] of the total observed variance to the remaining three measurable artifacts.

The remaining 25% of the observed variance is regarded as being caused by unmeasured statistical artifacts according to VG logic. Remember, a VG ratio = .75 implies $\underline{\sigma}_{\underline{p}}^2 = 0$ because 25% of the variance in $\underline{s}_{\underline{r}}^2$ can be attributed to unmeasured artifacts (Schmidt et al., 1982, p. 840). Unmeasured artifacts involve (a) between-study differences in criterion contamination and deficiency; (b) clerical errors in computation, typing, and transcription; and (c) "slight differences in the factor structure of tests designed to measure the same construct" (Schmidt et al., 1982, p. 840).

We find it incongruous that the variance attributed to criterion contamination and deficiency, clerical errors, and "slight" differences in test factor structures exceeds the variance attributed to range restriction, criterion reliability, and test reliability by a factor greater than 3 (i.e., .25/.075 = 3.33 at the critical value of the decision rule). The adjective "slight" in describing differences in factor structures of tests is well-taken, for if factor structures of tests vary among studies, then the VG analysis is mixing apples with bicycles. But how does one operationalize "slight" in regard to a point-estimate of variance attributable to this artifact? Well, a reasonable heuristic might be to interpret "slight" to mean approximately 10% of all of the variance attributed to unmeasured artifacts, or .10(.25) 100 = 2.5%. This suggests also that 2.5% of the total variance among the $\underline{r}_i$ (i.e., $\underline{s}_r^2$) could be attributed to "slight" differences in factor structures of tests, which if anything, seems generous given that only 7.5% of this variance is attributed to between-study differences in criterion reliability (CR), predictor reliabilty (PR), and range restriction (RR).

The rationale above means that approximately 22.5% of $\underline{s}_r^2$ should be viewed as being caused by the unmeasured artifacts of criterion contamination and deficiency (CCD) and clerical errors (CE). But is all of the variance among the $\underline{r}_i$ to be attributed to CCD and CE unmeasured? We think not because many of the causes of CCD and CE that affect between-study difference in validities are also likely to influence between-study differences in reliabilities. The logical and statistical progression is that at least some of the causes of $\underline{s}_r^2$ attributed to CCE and CE are in truth already measured and included in variance among the $\underline{r}_i$

attributed to CR and PR.

To illustrate, criterion contamination involves (a) systematic biases evolving from factors such as opportunity bias, rater bias, group characteristic bias, and knowledge of predictor bias; and (b) random error (Blum & Naylor, 1968). Between-study differences in random errors are obviously included in variation among the r_i attributed to between-study differences in CR. Variation in systematic biases over studies also influences variations in CR (cf. Guion, 1965; James, Demaree, & Wolf, 1984) and therefore variation among the r_i. Much the same can be said for clerical errors. Be these errors systematic and/or nonsystematic and involved in criterion and/or predictor measurement, they should influence variation in validities via variation in CR and PR, respectively.

In sum, it is our belief that variation among the r_i attributed to the unmeasured artifacts of CCD and CE is at least in part represented in the measured artifacts of CR and PR. This suggests (to us at least) that a VG decision rule which proportions roughly 22.5% of s_r^2 to the unmeasured artifacts of CCD and CD is seriously flawed, given that (a) a significant and likely substantial portion of the variance in validities attributed to CCD and CE is already represented in the measured artifacts of CR and PR, and (b) CR and PR, plus RR, account for only 7.5% of s_r^2 at the critical value of the decision rule. We propose, therefore, that a more reasonable hypothesis is that variance among the r_i due to truly unmeasured portions of the CCD and CE artifacts is unlikely to be greater than variance among the r_i that is caused by the measured artifacts CR, PR, and

RR--that is, 7.5% of $\underline{s}_{\underline{r}}^2$. So, if we estimate variance among the $\underline{r}_{\underline{i}}$ due to truly unmeasured sources in CCD and CE at 7.5%, add to this variance in the $\underline{r}_{\underline{i}}$ due to "slight" differences in the factor structures of tests (i.e., 2.5%), we have approximately 10% of $\underline{s}_{\underline{r}}^2$ attributable to unmeasured artifacts.

Proceeding on this basis suggests that given a VG ratio = .75, we should add .10 to the ratio to account for unmeasured artifacts (i.e., attribute 85% of the observed variance to artifacts). This leaves 15% of the variance attributable to $\underline{\sigma}_{\underline{p}}^2$. To reduce this value to zero--that is, to define a VG decision rule that more realistically implies that $\underline{\sigma}_{\underline{p}}^2 = 0$ --requires that we add .15 to .75. Thus, it is our recommendation that a VG decision rule of .90 should replace the present decision rule of .75. Naturally, this rule should be revised as research accumulates regarding empirical estimates of independent variance due to CCD, CE, and factor structures of tests. On the other hand, to leave the VG decision rule at .75 is to invite rejection of the null hypothesis that $\underline{\sigma}_{\underline{p}}^2 > 0$ when, according to the heuristics above, $\underline{\sigma}_{\underline{p}}^2$ could account for approximately 15% of the observed variance.

Summary and conclusions. The primary conclusions based on the preceding discussion are (a) VG procedures do not furnish irrefutable evidence of cross-situational consistency, and to imply that they do is to commit the logical fallacy of affirming the consequent; and (b) the VG decision rule of $\hat{\underline{\sigma}}_{\underline{e}}^2/\underline{s}_{\underline{r}}^2 \geq .75$ should be replaced with $\hat{\underline{\sigma}}_{\underline{e}}^2/\underline{s}_{\underline{r}}^2 \geq .90$, given that samples are not large and that the measurable artifacts are sampling error, range restriction, criterion reliabilty, and predictor reliability. Adopting a decision rule of .90 should reduce conflicts between the results of different types of analyses, such as between VG analysis and chi-square tests of the homogeneity of population correlations. Applied to the data in Table 1, for example, a decision rule of .90 would fail to disconfirm the null hypothesis that $\sigma_{\underline{p}}^2 > 0$, thus leaving $\sigma_{\underline{p}}^2 > 0$, which was confirmed by the chi-square analysis, as the most useful explanation for the observed variation among the $\underline{r}_{\underline{i}}$.

This example above is illustrative of the fact that a decision rule of .90 will likely reduce the percentage of occasions on which an inference that $\sigma_{\underline{p}}^2 = 0$ is warranted from the present 46% of validity distributions (Schmidt et al., 1982) to a lower, perhaps much lower, percentage of validity distributions. It follows that the hypothesis of situational specificity is alive and well (was it ever not?). However, the recommended change to a decision rule of .90 may stimulate the cry that (a) bias in favor of a finding of cross-situational consistency is being replaced with bias in favor of situational specificity, and/or (b) one heuristic decision rule (VG ratio $\geq$ .75) which lacks corroborative evidence is merely being replaced with another heuristic decision rule (VG ratio $\geq$ .90) that is equally lacking in corroborative evidence. In response, we largely agree with the latter point and again underscore the need for research designed to identify an empirically defensible decision rule. In the interim, we believe that a decision rule of .90 is more reasonable than a decision rule of .75 for the reasons stated in the development of the recommended change to a rule of .90.

Finally, we refer again to the Osburn et al. (1983) simulation study which clearly supported the need for a decision rule more stringent than .75.

False Assumptions

It was noted briefly that the statistical foundation for the VG analytic procedures is furnished by psychometric analogy and structural equations (cf. Hunter, Schmidt, & Jackson, 1982; Schmidt et al., 1980). The fundamental structural equation is $r_i = p_i + e_i$, where r_i is the observed score for sample (subject) i, p_i is the population correlation (true score) for sample (subject) i, and e_i is the sampling (random measurement) error for sample (subject) i. Like the psychometric equation on which it is based, this equation is underidentified. That is, for each sample we have one piece of known data (r_i) and two pieces of unknown data (p_i and e_i). We thus have one equation in two unknowns, the result of which is no unique mathematical solution for either unknown. Adding new samples from different populations does not help because each new sample contributes one known and two unknowns, not to mention a new population and a new sampling distribution. (This discussion and that to follow is based on the usual case of one sample per situation or population. If it were possible to obtain many independent, random samples from each situation, then not only could each p_i be estimated, but also the total variance among all observed correlations could be decomposed empirically into between-situation variance and within-situation variance, using classic ANOVA paradigms. Unfortunately, the rarity of many independent samples from each of two or more different situations requires that we proceed with but one of a theoretically infinite number of samples from each of K populations and sampling distributions.)[1]

Similar to classic psychometrics (cf. Gulliksen, 1950; Lord & Novick, 1968), the VG approach proceeds with the underidentified structural equation and employs a set of assumptions that make possible the estimation of moments of the unobservable (latent) true scores and error scores in terms of moments of the observed r_i. Given the basic structural equation $r_i = p_i + e_i$, it is assumed that (a) the mean error is zero within each study (population, situation), (b) p_i and e_i are unrelated across studies, and (c) $\sigma_r^2 = \sigma_p^2 + \sigma_e^2$ (Hunter, Schmidt, & Jackson, 1982). Furthermore, implicit in the use of several VG estimating equations is the assumption that the errors are normally distributed and/or the assumption that the within-study error variances are homogeneous. Each of these assumptions is discussed in greater detail below, where it is shown that all of the assumptions above are false if the p_i vary or could vary.

Nonnormality of error distributions. A theoretical sampling distribution of observed correlations (r_{ia}) exists for each p_i, where i again references 1, ..., K populations and a refers to 1, ..., A observed correlations in the sampling distribution for each p_i (technically, $A \rightarrow \infty$). The variance among the r_{ia} for a particular p_i is designated σ_{ri}^2. Given that $r_{ia} = p_i + e_{ia}$ in a given sampling distribution and that p_i is a constant in that population, it follows that $\sigma_{ri}^2 = \sigma_{ei}^2$, where σ_{ei}^2 is the error variance for the sampling distribution associated with p_i. (Note that σ_r^2 and σ_e^2 refer to variances *over K populations*.) The equation employed in VG analysis to estimate error variance for a sampling distribution derives from the equation:

$$\sigma_{ei}^2 = \sigma_{ri}^2 = (1-p_i^2)^2/n \tag{6}$$

which assumes a large sample (see below) drawn from a bivariate normal population with correlation coefficient p_i (Kendall & Stuart, 1969, 1973).

The sample estimating equation based on the single r_{ia}, or r_i, used in VG analysis is:

$$\hat{\sigma}_{ei}^2 = s_{ri}^2 = (1-r_i^2)^2/(n_i-1) \tag{7}$$

which is presented and discussed by Ezekiel and Fox (1959) and Fisher (1954) (Fisher also uses (n_i -1) in the denominator of Equation 6).

Kendall and Stuart (1973, p. 304) contended that the use of Equation 6 (and by implication Equation 7) to estimate the variance of a sampling distribution "is of little value in practice since the distribution of r tends to normality so slowly [cf. Kendall & Stuart, 1969, p. 388]: it is unwise to use it for [n_i] < 500." Fisher (1954) suggested that Equation 7 should not be used for an n_i < 100. The rationale for these statements is that when p_i departs from zero and n_i is not large (e.g., < 100 or 500, depending on the reference), then the distribution of the r_{ia} is skewed. In particular, the distribution is negatively skewed for positive p_i. For a given n_i, such as 70, the degree of skew, as well as kurtosis, increases as the (absolute) value of p_i increases (Ezekiel & Fox, 1959; Fisher, 1954; Kendall & Stuart, 1969, 1973; Muirhead, 1982). In general, the sampling distribution tends toward normality, but very slowly, as n_i increases, although with very large p_i the distribution remains nonnormal even with large n_i.

Focusing on positive p_i, the reason for negative skews is simple; p_i is bounded by 1.00. The problem is therefore most pronounced for very large p_i. Nevertheless, even with moderate p_i and $n_i <$ 100 or 500, a ramification of negative skews in sampling distributions for most if not all of the correlations in Table 1 is that the estimate of error variance for each of the K samples is less than it should be (cf. Fisher, 1954). It follows that the estimate of expected error variance furnished by Equation 3 in Table 2 is also less than it should be. We can correct these estimates by using the asymptotic expansion furnished by Ghosh (1966) for estimating the variance of the r_{ia} for a single sampling distribution. Unfortunately, this equation is too complex to present here. In general, however, the values furnished by the Ghosh (1966) equation for the correlations in Table 1 are only slightly larger than those furnished by Equation 7, given $n_i = 70$. For example, with $r_i = .50$, $\hat{\sigma}_{ei}^2$ is .0083 based on the Ghosh (1966) equation and .0082 based on Equation 7.

The difference between the correct estimates furnished by the Ghosh

(1966) equation and those furnished by Equation 7 are trivial (for these data), and one may argue that the practical approach is to employ Equation 7 to estimate error variance. But a plea for pragmatism (and expediency) is confronted with the problem that VG procedures were developed to explain why observed validities vary over situations by testing causal models for observed validities and distributions of observed validities. Indeed, causal models and structural equations are presumably employed in VG analysis in order to "greatly advance" the scientific status of personnel research. But scientific explanation is not "greatly advanced" by relying on an equation (Equation 7) that statisticians have shown to be flawed, however trivial the flaw, for smal' samples and $\underline{p}_i \neq 0$, the key constituents of VG analysis.

Yet an important commodity in science is time, and the time and difficulty required to use the Ghosh (1966) equation versus Equation 7 are compelling forces to proceed with the pragmatic, indeed parsimonious, use of Equation 7 to estimate error variance. However, a call for parsimony and pragmatism is not a defensible position in this case because it is unnecessary. Specifically, a minimal amount of time spent in converting the $\underline{r}_i$ to Fisher $\underline{z}$ coefficients ($\underline{z}$s) would help to resolve not only the problem of nonnormal distributions -- distributions of $\underline{z}$s approach normality much more rapidly than (Pearson) $\underline{r}$s -- but also most of the statistical errors discussed below. Thus, we do not urge the use of the Ghosh (1966) equation or any other equation based on correlations. We will recommend the use of $\underline{z}$s in VG analysis. Before developing these points, however, it is necessary to document other problems with the use of rs in VG analysis.

Nonzero expected values of errors. Hunter, Schmidt, and Jackson (1982, p. 43) state that "Since the mean error is zero within each study, the error variances across studies in [sic] the average within study variance." The first problem with this statement is that the mean within-study (within-population) error is not equal to zero with skewed distributions. This point derives from the well-known fact that an r_{ia} is a biased indicator of its respective p_i (cf. Muirhead, 1982). The expected value for e_{ia} is:

$$\begin{aligned} E(e_{ia}) &= E(r_{ia}) - p_i \\ &= (p_i - [\, p_i(1-p_i{}^2)]/2[n_i - 1]) - p_i \\ &= [\, -p_i(1-p_i{}^2)/2(n_i - 1)] \qquad (8) \end{aligned}$$

This derivation is based on Muirhead's (1982) equation for $E(r)$ and involves deletion of a term $O(n^{-2})$ from the $E(r)$ equation. Equation 8 suggests that the mean error within each study takes a negative value for positive p_i, which is expected for negatively skewed sampling distributions. It suggests also that if the p_i vary, then the $E(e_{ia})$ will also vary because $E(e_{ia})$ is a function of p_i. This connotes that some variation among the r_i *over studies* could be due to variation among the means of the within-study errors. Finally, given that the $E(e_{ia})$ are a function of the p_i, the possibility exists that the e_i and p_i in the equation $r_i = p_i + e_i$

are related ($\underline{A}$=1 in this equation). We cannot show this directly with our illustrative data because the $\underline{p}_i$ are unknown (i.e., we have only reasonable limits). We may, however, develop another illustration.

A hypothetical distribution of 14 $\underline{p}_i$ (true validities) is presented in Table 3. The $\underline{p}_i$ vary between .05 and .70. Values of $\underline{E}(\underline{e}_{ia})$ are given for each $\underline{p}_i$; these values are based on $\underline{n}_i$ = 70 for all samples. The values of the $\underline{E}(\underline{e}_{ia})$ are of small magnitude, which suggests minimal bias in variance estimates because of failure to consider variation in the expected errors. More important is the curvilinear relation between the $\underline{p}_i$ and the $\underline{E}(\underline{e}_{ia})$. Technically, $\underline{E}(\underline{e}_{ia})$ assumes a maximum value at $\underline{p}_i$ =.58, approximately. As $\underline{p}_i$ increases from .05 to .58, the values of the $\underline{E}(\underline{e}_{ia})$ become increasingly negative; as $\underline{p}_i$ increases beyond .58, the values of the $\underline{E}(\underline{e}_{ia})$ become decreasingly negative. Inasmuch as the true validity for a single test would not be expected to exceed .58 very often, we might assume that the $\underline{p}_i$ and $\underline{E}(\underline{e}_{ia})$ are generally negatively related. We pursue this point below.

Insert Table 3 about here

Nonhomogeneous error variances. The equation for the variance of errors for a sampling distribution (Equation 6) shows that $\underline{\sigma}_{ei}^2$ varies as function of $\underline{p}_i$. For positive $\underline{p}_i$, $\underline{\sigma}_{ei}^2$ is inversely related to the magnitude of $\underline{p}_i$. This results in violation of the usual assumption in psychometrics that error variances associated with different true scores are homogeneous. A more important consideration, however, is the inverse relation between the $\underline{p}_i$ and the $\underline{\sigma}_{ei}^2$ and the implications of this relation for independence between the the $\underline{p}_i$ and the $\underline{e}_i$.

Nonindependence between true scores and error scores. Two critical equations in VG analysis, namely the VG ratio $\hat{\underline{\sigma}}_e^2/\underline{s}_r^2$ and "est $\underline{\sigma}_p^2 = \underline{\sigma}_r^2 - \underline{\sigma}_e^2$" (Hunter, Schmidt & Jackson, 1982, p. 44), are contingent on the assumption that $\underline{\sigma}_r^2 = \underline{\sigma}_p^2 + \underline{\sigma}_e^2$, which in turn is based on the assumptions that the $\underline{p}_i$ and $\underline{e}_i$ are independent (over $\underline{K}$ populations or studies) and that $\underline{E}(\underline{r}_i) = \underline{p}_i$. It was noted that generally $\underline{E}(\underline{r}_i) \neq \underline{p}_i$ and thus we now turn to the assumption that $\underline{p}_i$ and $\underline{e}_i$ are independent. Lord (1960, p. 94) referred to the assumption of independence between true scores and error scores as the "independence hypothesis." Lord (1960, p. 91) also noted that the key concern is the "hypothetical bivariate scatterplot between true scores and errors of measurement," which "cannot be constructed empirically." A similar rationale applies here; we wish to know the relation between the $\underline{p}_i$ and the $\underline{e}_i$. We cannot construct a bivariate scatterplot because we do not know the values of either the $\underline{p}_i$ or the $\underline{e}_i$ inasmuch as the equation $\underline{r}_i = \underline{p}_i + \underline{e}_i$ is underidentified. We may, however, address lack of independence by other means. For example, a hypothetical set of $\underline{p}_i$ and $\underline{E}(\underline{e}_{ia})$ values in Table 3 implied nonindependence between the $\underline{p}_i$ and $\underline{e}_i$. This issue is now treated using procedures furnished by Lord (1960) and Lord and Novick (1968).

These authors recommended the use of third-order moments to test the independence hypothesis. The test of concern is based on the covariance between the p_i (in deviation form) and conditional error variances (i.e., the σ_{ei}^2), or Cov(p_i σ_{ei}^2) (Lord & Novick, 1968, p. 229). If the p_i are independent of the e_i, then Cov (p_i σ_{ei}^2) = 0. But this is obviously not the case because, as discussed, not only are the σ_{ei}^2 nonhomogeneous, but also the σ_{ei}^2 vary inversely as a function of the p_i. Thus, for positive p_i, Cov(p_i σ_{ei}^2) assumes a nonzero, negative value, from which we can conclude that the p_i and e_i are nonindependent. It follows that the equation $\sigma_r^2 = \sigma_p^2 + \sigma_e^2$ is in error and that the statistical foundation for such things as the VG ratio is also in error.

Satisfying the assumptions. The attempt to use the classic psychometric model to build a statistical foundation for VG analysis results in violation of many of the classic model's assumptions. Now, if $p_i = p$ (i.e., σ_p^2=0), as the VG proponents assume, then many of the problems discussed in regard to false assumptions dissolve. For example, the error variances are homogeneous because all p_i are the same (for constant n_i). However, assuming $p > 0$, the sampling distributions are still likely to be negatively skewed, the expected error variances are not zero, and the error variance is underestimated using the VG equation. The primary result of these problems is a small bias in favor of a finding of situational specificity. Yet, the VG procedures remain troublesome because they are represented as a test of the hypothesis that σ_p^2 >0.

But the moment we necessarily entertain the possibility that the p_i vary, we must also allow for the possibility of heterogeneous error variances and nonindependence between the p_i and e_i. Thus, under the stated basis for the null hypothesis, and making the not unrealistic assumption that the p_i are never precisely equal, it follows directly that the VG procedures furnish biased estimates because statistical assumptions are not satisfied.

How serious is the bias? An example presented shortly indicates a small overall bias in favor of a finding of cross-situational consistency for the illustrative data in Table 1. Other investigators have addressed at least some of the assumptions for VG analysis and have concluded that (a) sampling distributions of observed correlations are "approximately normal" (Pearlman et al., 1980 p. 381) or "close to normal" (Schmidt et al., 1981, p. 174) except for very large values of p, which implies little or nc bias due to skewed sampling distributions for selection studies at least; and (b) nonindependence between the p_i and e_i results in minor underestimation of the value of σ_p^2 (Burke, 1984; Linn & Dunbar, 1982). Callender, Osburn, Greener, and Ashworth (1982) used Monte Carlo techniques to show that a skewed distribution of hypothetical p_i had no influence on estimates of σ_p^2. The general conclusion, therefore, appears to be that whatever bias exists in VG analysis is small and, pragmatically, has little influence on results. Consequently, one may proceed with VG analysis without grave concern for bias introduced by violations of assumptions.

A reasonable opinion, but not one that we share. To reiterate briefly, our view is that if (a) an avowed objective of using VG procedures is to advance the scientific status of personnel research (cf. Schmidt & Hunter, 1977), then (b) the VG procedures should stand up to scientific scrutiny. While science is not blind to the need for pragmatics and occasional expediency, the pattern of formal statistical error after formal statistical error does little to advance the scientific merits of the VG enterprise. Even if the overall degree of bias is trivial, it would hardly do to attempt to promote the VG procedures as a scientific advancement when pragmatics are the justifications for violations of almost every statistical assumption of the model. Most importantly, it is unnecessary to have to rely to this degree on pragmatics because a simple solution exists that reduces the bias and increases the scientific precision of the VG procedure.

As noted briefly, the simple solution is to transform the observed validities (r_i) into Fisher z coefficients and to base the VG analysis on these coefficients. For sample sizes greater than 50, the sampling distribution of zs is approximately normal, irrespective of the value of p_i (Kendall & Stuart, 1969, who also present estimation equations for $n_i \leq 50$). Furthermore, σ_{ei}^2 based on zs is essentially independent of the value of p_i because all σ_{ei}^2 have an estimated value of $1/(n_i-3)$ (for constant n_i; variable n_i is addressed by weighting in VG analysis). A very slight bias may persist if $E(e_{ia})$ based on z coefficients is not zero, but this is an approximation that we can live with (see Hotelling, 1953 and Kendall & Stuart, 1969 for further discussions of this issue).

In any event, the use of z coefficients in place of correlation coefficients places VG analysis on a sounder statistical footing even though it does not ameliorate all of the statistical problems. Interestingly, Schmidt and Hunter (1977) originally used *z*s in VG analysis to ensure against covariation between the p_i and e_i, but switched to observed correlations under the assumption that their formula for sampling error was "very accurate" (Schmidt et al., 1980, p. 660). Later, the reason for the switch from *z*s to *r*s was given as "the effect of Fisher's *z* transformation is to assign extra weight to large observed validity coefficients" (Schmidt et al., 1982, p. 839). We interpret the term "assign extra weight" to mean that the difference between the value of *z* and the value of *r* increases in absolute value as the value of *r* increases. This, of course, is the price one pays to achieve a sampling distribution of *z*s that approaches normality more quickly than a sampling distribution of *r*s. It also suggests that the variance among the *z*s will be greater than the variance among the *r*s and that the VG ratio will tend to be lower for *z*s than for *r*s. These points are illustrated by a reanalysis of the data in Table 1 using Fisher *z* transformations. The VG ratio based on *z*s is .0149/.021 = .71, where $\hat{\sigma}_e^2 = 1/67$. The mean Fisher *z* for the transformed r_i in Table 1 is .56, and the variance of the *z*s is .021. Clearly, .71 differs little from the VG ratio of .75 based on *r*s and Equation 4 (Table 2), although the difference does indicate a slight bias in VG procedures based on *r*s in favor of a finding of cross-situational consistency.

It is important to note that this slight bias overestimates the bias that would likely be obtained with real selection data because the illustrative data in Table 1 involve only sampling error and variation about a mean correlation (i.e., .50) equal to what is considered by VG proponents to be the "true validity" of many tests (cf. Pearlman et al., 1980). With real data the VG approach works with a distribution of correlations whose values have also been attenuated and/or reduced by criterion (and predictor) measurement error and range restriction in estimating the predicted variance among the observed correlations that is due to measurable statistical artifacts (cf. Pearlman et al., 1980; Schmidt et al., 1980). The corollaries to this point are that (a) the absolute magnitudes of the r_i will be lower than those in Table 1, from which it follows that (b) the difference between statistical values based on r and z, such as the VG ratio, will be reduced because the differences between values of r and values of z decrease as the absolute value of r decreases. But then VG techniques are not limited to selection research and may be employed for distributions in which the rs are of greater magnitude than typically found in selection studies or the illustration used here. Indeed, VG analysis based on rs may be inappropriate as a general method for its flaws become increasingly evident as the correlations increase in magnitude and assumptions become more tenuous. Of course, a simple and statistically more precise alternative exists, namely to use z coefficients in analyses.

Recommendations and Conclusions

Three major recommendations have been proposed. First, inferences based on the results of VG research should be less dramatic. Empirical support for a cross-situational consistency model implies only that this model furnishes a useful basis for explaining the distribution of observed validities. This model is not unique, irrefutable, or proven. Second, the decision rule for the VG ratio should be .90 rather than .75. This recommendation indicates that 90% of the variance in observed validities should be attributed to sampling error and other measurable artifacts before one infers that $\sigma_{\underline{p}}^2 = 0$. The remaining 10% of the variance is assumed to be due to unmeasured artifacts. This recommendation is subject to immediate change as soon as research is obtained pertaining to the unique influences of criterion problems, clerical errors, and predictor factor structures on variation among validities. Third, VG analyses should employ Fisher $\underline{z}$ coefficients rather than (Pearson) correlation coefficients. The objective here is to place VG analysis on a sounder statistical footing.

A likely result of the second and third recommendations, especially the proposed change in the VG decision rule (which applies to the use of Fisher $\underline{z}$ coefficients in VG analysis), is that fewer VG analyses will conclude that cross-situational consistency is a useful model for explaining variation in validities. This conclusion has the somewhat unfortunate implication that all validities must therefore vary. There are other views. Heretofore we have focused on extremes for the purpose of contrast. Now let us ask whether it is realistic to assume that $\underline{p}$ is different for each situation, or at least different enough to warrant a separate analysis for each situation? Probably not. But it is, in our opinion, as realistic as assuming that $\underline{p}$ is a constant for every situation or, at least, that the $\underline{p}$'s do not vary sufficiently to warrant separate analyses for at least some situations. Fortunately, situational specificity and cross-situational consistency as described in this paper are but two of many possible views. Indeed, the most useful models for explaining variation in validities probably lie in some middle ground between these two extremes. This is not the place to attempt to review a voluminous literature on theoretical models. We suggest only that attempts to assess situational specificity and cross-situational consistency will be enhanced by including situational variables in analyses. Measures representing membership in gross categories such as job families (cf. Pearlman, 1980) are helpful but lack the explanatory power furnished by measurement and explicit analysis of specific aspects of situations (e.g., stress, leadership) that presumably influence correlations between person variables and job performance (cf. James, Demaree, & Hater, 1980). An ideal strategy would be to attempt to develop structural (causal, explanatory) models of job performance (and attitudes) that involve both person variables and situational variables.

In closing, although we have been critical of VG procedures, we do believe that the VG approach is creative and has the potential to make a contribution to research. Our key concern has been the overdramatic interpretations of the results of validity generalization analyses in favor of cross-situational consistency. These concerns apply also to validity generalization to the extent that "estimated true validities" and "credibility values" are subject to alternative models involving a potentially greater degree of situational specificity than indicated by VG assumptions and decisions rules. On the other hand, the issue of differential validity in the context of validity generalization (cf. Hunter & Hunter, 1984) is outside the bounds of this discussion. Treatment of systematic variation <u>within</u> <u>situations</u> due to such things as ethnic, race, and sex distinctions requires additional thought and statistical modeling. Finally, as indicated by the preceeding discussion, no attempt was made to exhuast all possible concerns with VG procedures. The issues addressed here were selected because they were considered to be among the more salient issues at this time, especially in the context of testing the goodness of fit of causal models for validities and validity distributions.

References

Blum, M.L. & Naylor, J.C. (1968) Industrial Psychology: It's Theoretical and Social Foundation. New York: Harper and Row

Bowers, K.S. (1973) Situationism in psychology: An analysis and a critique. Psychological Review, 80, 307-336.

Burke, M.J. (1984) Validity generalization: A review and critique of the correlation model. Personnel Psychology, 37, 93-113.

Callender, J.C., & Osburn, H.G. (1980) Development and test of a new model for validity generalization. Journal of Applied Psychology, 65, 543-558.

Callender, J.C., & Osburn, H.G. (1981) Testing the constancy of validity with computer generated sampling distributions of the multiplicative model variance estimate: Result for petroleum industry validation research. Journal of Applied Psychology, 66, 274-281.

Callender, J.C., & Osburn, H.G. (1982) Another view of progress in validity generalization: Reply to Schmidt, Hunter, and Pearlman. Journal of Applied Psychhology, 67, 846-852.

Callender, J.C., Osburn, H.G., Greener, J.M., & Ashworth, S. (1982) The multiplicative validity generalization model: Accuracy of estimates as a function of sample size and mean, variance and shape of the distribution of true validities. Journal of Applied Psychology, 67, 859-867.

Cohen, J., & Cohen, P. (1975) Applied multiple regression/correlation analysis for the behavioral sciences. New York: Wiley.

Cohen, J., & Cohen, P. (1983) Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.), Hillsdale, NJ: Erlbaum.

Ekehammer, B. (1974) Interactionism in personality from a historical perspective. Psychological Bulletin, 81, 1026-1048.

Endler, N.S. & Magnusson, D. (1976) Toward an interactional psychology of personality. Psychological Bulletin,83, 956-974.

Ezekiel, M., & Fox, K.A. (1959) Methods of correlation and regression analysis. New York: Wiley.

Fisher, R.A. (1954) Statistical methods for research workers. Edinburgh: Oliver and Boyd.

Ghiselli, E.E. (1966) The validity of occupational aptitude tests. New York: Wiley.

Ghiselli, E.E. (1973) The validity of aptitude tests in personnel selection. Personnel Psychology, 216, 461-477.

Ghosh, B.K. (1966) Asymptotic expansions for the moments of the distribution of correlation coefficient Biometrika, 53, 258-262.

Guion, R.M. (1965) Personal testing. New York: McGraw-Hill.

Gulliksen, H. (1950) Theory of mental tests. New York: Wiley.

Hunter, J.E. & Hunter, R.F. (1984) Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.

Hunter, J.E., Schmidt, F.L., & Jackson, G.B. (1982) Advanced meta-analysis: Quantitative methods for cumulating research findings across studies. Beverly Hills, CA: Sage.

Hunter, J.E., Schmidt, F.L., & Pearlman, K. (1982) The history and accuracy of validity generalization equations: A response to the Callender and Osburn reply. Journal of Applied Psychology, 67, 853-858.

James, L.R., Demaree, R.G., & Hater, J.J. (1980) A statistical rationale for relating situational variables and individual differences. Organizational Behavior and Human Performance, 25, 354-364.

James, L.R., Demaree, R.G., & Wolf, G. (1984) Estimating within-group interrater reliability with and without response bias. Journal of Applied Psychology 69, 85-98.

James, L.R., & Jones, A.P. (1976) Organizational structure: A review of structural dimensions and their conceptual relationship with individual attitudes and behavior. Organizational Behavior and Human Performance, 16, 74-113.

James, L.R., Mulaik, S.A., & Brett, J.M. (1982) Conditions for confirmatory analysis and causal inference. Beverly Hills: Sage.

Kendall, M.G., & Stuart, A. (1969) The advanced theory of statistics: Volume 1. London, England: Griffin.

Kendall, M.G., & Stuart, A. (1973) The advanced theory of statistics: Volume 2. London, England: Griffin.

Lent, R.H., Aurbach, H.A., & Levin, L.S. (1971) Research design and validity assessment. Personnel Psychology, 24, 247-274.

Lewin, K. (1938) The conceptual representation of the measurement of psychological forces. Durham, N.C.: Duke University Press.

Lichtman, C.M., & Hunt, R.G. (1971) Personality and organization theory: A review of some conceptual literature. Psychological Bulletin, 76, 271-294.

Linn, R.L., & Dunbar, S.B. (1982, November) Validity generalization and predictive bias. Paper presented at the Fourth Johns Hopkins University National Symposium on Educational Research, Washington, D.C.

Lord, F.M. (1960) An empirical study of the normality and independence of errors of measurement in test scores. Psychometrika, 25, 91-104.

Lord, F.M., & Novick, M.R. (1968) Statistical theories of mental test scores. Reading, M.A.: Addison-Wesley.

Muirhead, R.J. (1982) Aspects of multivariate statistical theory. New York: Wiley.

Osburn, H.G., Callender, J.C., Greener, J.M., & Ashworth, S. (1983) Statistical power of tests of the situational specificity hypothesis in validity generalization studies: A cautionary note. Journal of Applied Psychology, 68, 115-122.

Pearlman, K. (1980) Job families: A review and discussion of their implications of personnel selection. Psychological Bulletin, 87, 1-28.

Pearlman, K., Schmidt, F.L., & Hunter, J.E. (1980) Validity generalization results for tests used to predict training success and job proficiency in clerical occupations. Journal of Applied Psychology, 65, 373-406.

Pervin, L. (1968) Performance and satisfaction as a function of individual environment fit. Psychological bulletin. 69, 56-58.

Raju, N.S., & Burke, M.J (1983) Two new procedures for studying validity generalization. Journal of Applied Psychology, 68, 382-395.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J.E. (1980) Validity generalization results for computer programmers. Journal of Applied Psychology, 65, 643-661.

Schmidt, F.L., & Hunter, J.E. (1977) Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Schmidt, F.L., & Hunter, J.E. (1978) Moderator research and the law of small numbers. Personnel Psychology, 31, 215-232.

Schmidt, F.L., & Hunter, J.E. (1980) The future of Criterion-related validity. Personnel Psychology. 33, 41-60.

Schmidt, F.L., & Hunter, J. E. (1981) Employment testing: Old theories and new research findings. American Psychologist, 36, 1128-1137.

Schmidt, F.L., & Hunter, J.E. (1984) A within setting empirical test of the situational specificity hypothesis in personnel selection. Personnel Psychology, 37, 317-326.

Schmidt, F.L., Hunter, J. E., & Pearlman, K. (1981) Task differences and validity of aptitude tests in selection: A red herring. Journal of Applied Psychology, 66, 166-185.

Schmidt, F.L., Hunter, J.E., & Pearlman, K. (1982) Progress in validity generalization: Comments on Callender and Osburn and further developments. Journal of Applied Psychology, 67, 835-845.

Schmidt, F.L., Hunter, J.E., Pearlman, K., & Hirsh, H.R. (1984) Questions and answers about validity generalization and meta-analysis. Unpublished manuscript.

Schmidt, F.L., Hunter, J.E., Pearlman, K., & Shane, G.S. (1979) Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. Personnel Psychology, 32, 257-281.

Authors' Notes

Support for this project was provided under Office of Naval Research Contract N00014-83-K-0480, Office of Naval Research Project NR475-026, which is part of the Manpower R and D Program. Opinions expressed are those of the authors and are not to be construed as necessarily reflecting the official view or endorsement of the Department of the Navy.

The authors wish to thank Jeanne M. Brett, Philip Bobko, Jack M. Feldman, Sigrid B. Gustafson, Raymond A. Katzell, Charles E. Lance, Edward E. Loveland, John F. White, and C. Michael York for their helpful suggestions and advice.

Please address requests for reprints to Lawrence R. James, School of Psychology, Georgia Institute of Atlanta, Atlanta, Georgia 30332.

Footnotes

[1] A recent study by Schmidt and Hunter (1984) applied VG procedures to validity data obtained from four different cohorts of stenographers ($\underline{K} = 4$) from the same organization. The objective of the study was to demonstrate that "if the statistical artifacts operating are the same, observed validities will vary as much within the same setting as they do across settings . . . merely as a result of artifacts such as sampling error" (Schmidt & Hunter, 1984, p. 320). Unfortunately, a sample of only four correlations was available for each of five different tests. The instability of results based on such a small sample of $\underline{r}$s is indicated by the values of the VG ratio (i.e., $\hat{\sigma}_{\underline{e}}^2/\underline{s}_{\underline{r}}^2$) for the five tests, which were 4.0, 1.31, .709, .422, and .81. (Schmidt and Hunter [1984] reported results in terms of a "new" ratio, namely $\underline{s}_{\underline{r}}/\hat{\sigma}_{\underline{e}}$, which took values of .50, .88, 1.19, 1.54, and 1.11 for the five tests, respectively). One should probably question the generalizability of data which, based on the "old" VG ratio, suggest for one test that 400% of the observed variance for correlations is accounted for by sampling error. Thus, we will not address this study again in this paper.

Table 1

Contrived Validity Coefficients

.72	.54	.45
.68	.53	.44
,65	.52	.43
.63	.51	.42
.61	.50	.41
.59	.50	.39
.58	.49	.37
.57	.48	.35
.56	.47	.32
.55	.46	.26

Table 2

Validity Generalization Equations and Estimates

1. Mean observed validity coefficient ($\bar{r}$):

$$\bar{r} = \frac{\sum_{K} r_i n_i}{\sum_{K} n_i} = \frac{\sum_{K} r_i}{K} = .50$$

where all n_i = 70 and K = 30 studies or situations

2. Variance among observed validity coefficients (s_r^2):

$$s_r^2 = \frac{\sum_{K} n_i (r_i - \bar{r})^2}{\sum_{K} n_i} = \frac{\sum_{K} (r_i - \bar{r})^2}{K} = .011$$

3. Estimate of the variance among observed validity coefficients expected due to sampling error (≐ average within study variance)($\hat{\sigma}_e^2$):

$$\hat{\sigma}_e^2 = \frac{\sum_{K} [n_i (1 - r_i^2)^2 / (n_i - 1)]}{\sum_{K} n_i} \doteq \frac{(1 - \bar{r}^2)^2}{(n_i - 1)} = .0082$$

4. Estimate of the proportion of variance in observed validity coefficients (s_r^2) that is attributable to sampling error ($\hat{\sigma}_e^2$):

$$\hat{\sigma}_e^2 / s_r^2 = .75$$

Table 3

Population Correlations and Expected Values of Sampling Errors

Population Correlation ($\underline{p}_{\underline{i}}$)	Expected Value of Error [$\underline{E}(\underline{e}_{\underline{ia}})$]
.70	-.00259
.65	-.00272
.60	-.00278
.55	-.00278
.50	-.00272
.45	-.00260
.40	-.00243
.35	-.00223
.30	-.00198
.25	-.00170
.20	-.00139
.15	-.00106
.10	-.00072
.05	-.00036

Note. $\underline{n}_{\underline{i}}$ = 70 for all samples.